Journal of
Health & Medical Informatics

Hailemariam et al., J Health Med Informat 2015, 6:4
DOI: 10.4172/2157-7420.1000197

Research Article Open Access

Application of Data Mining Techniques to Predict Adult Mortality: The Case of Butajira Rural Health Program, Butajira, Ethiopia

Tesfahun Hailemariam[1*], Million Meshesha[2] and Alemayehu Worku[3]

[1]Department of Health Informatics, Hawassa Health Science College, Hawassa, Ethiopia
[2]Department of Information Science, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
[3]School of Community Health Department, Addis Ababa University, Addis Ababa, Ethiopia

Abstract

Background: Though adults are care providers and risk takers of a society, reports indicate that adult mortality conditions are not given much emphasis. This is due to a widespread perception that mortality among adults is low. Every year, more than 7•7 million children die before their fifth birthday; however, nearly 24 million of adults die under the age of 70 years. Identifying major determinants for adult death helps to alleviate the loss of the productive group. Therefore, this research is aimed to apply data mining techniques to build a model that can assist in predicting adult health status.

Methods: The hybrid model that was developed for academic research was followed. Dataset was preprocessed for data transformation, missing values and outliers. WEKA 3.6.8 data mining tools and techniques such as J48 decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869 instances of both alive and died adults through three experiments and six scenarios. The area under the ROC curve for outcome class is used to evaluate performances of models from the predictive algorithms.

Results: In this study as compared to Bayes, the performance of J48 pruned decision tree reveals that 97.2% of accurate results are possible for developing classification rules that can be used for prediction. If no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death is 98.4% and 97.4% respectively with concomitant attributes in the rule generated. The likely chance of adult to survive in completed primary school, completed secondary school, and further education is (98.9%, 99%, 100%) respectively.

Conclusion: Predictive model built with the use of data mining techniques suggests that education plays a considerable role as a root cause of adult death, followed by outmigration. The possibility of incorporating the findings of this study with knowledge based system should be explored so that experts can consult the system in their problem solving and decision making process. Further comprehensive and extensive experimentation is needed to substantially describe the loss experience of adult mortality in Ethiopia.

Keywords: Predictive modeling; Adult mortality; Adult; Data mining; BRHP dataset

Introduction

Living long is a much desired aspiration by everyone. This is because living is not only as a state of being itself valued, but also it is a necessary requirement for carrying our plans that we have reason to value [1]. The reason for reversal in the adult life expectancy in Africa and some of the developing nations is due to premature adult death and the wide gap of ranges in developed and developing societies [2]. Adult mortality is the probability that a 15 years old person will die before reaching his/her 60th birth day or probability of dying between 15 to 60 years per 1000 population [3,4]. Though adults are care providers and risk takers of a society, reports indicate that adult mortality conditions are not given much emphasis. This is due to a widespread perception that mortality among adults is low [5]. Every year, more than 7·7 million children die before their fifth birthday; however, over three times of 7.7 million of adults i.e. nearly 24 million die under the age of 70 years [6]. The risk of a 15-year-old dying before reaching 60 years of age is 12% for men and 5% for women in developed countries where as the risk of dying is double in developing countries which is 25% and 22% for men and women respectively [5]. These gaps possibly can be narrowed through improving life expectancy of the adult through tackling likely causes of adult mortality.

MDG5 declaration on maternal health, focuses on one of the important causes of death in women aged 15–49 years to track its goal. In order to attain the millennium promises regarding to adult health, particularly maternal mortality ratio, and adult female mortality rates are pledged as an essential component of the measurement [6]. Pledges of emphasis on adult health is due to adult mortality has received little policy attention, resources and monitoring efforts. Understanding of this phenomenon has an important implication in terms of promotive and preventive aspect of health care which depend on population based analysis that needs timely, accurate, complete and adequate information on demographic characteristics and predisposing factors of health problems among the population [7]. Butajira Rural Health Programme (BRHP) has been established in 1986 as an epidemiological study that approaches to identify major variables with their relationships in order to provide up-to dated epidemiological information so as to improve health care provision, planning and decision-making at the district using statistical tools [7].

The fast-growing, tremendous amount of data which is collected

***Corresponding author:** Tesfahun Hailemariam, Department of Health Informatics, Hawassa Health Science College, Hawassa, Ethiopia, Tel: +251934107979; E-mail: tesfahunhailemariam@gmail.com

Received May 27, 2015; **Accepted** July 24, 2015; **Published** July 31, 2015

Citation: Hailemariam T, Meshesha M, Worku A (2015) Application of Data Mining Techniques to Predict Adult Mortality: The Case of Butajira Rural Health Program, Butajira, Ethiopia. J Health Med Informat 6: 197. doi:10.4172/2157-7420.1000197

and stored in large and numerous data repositories has far exceeded human ability for comprehension without powerful tools [8]. Knowledge from the bulk of data on trends, in age, gender, geographic variations, and burden of disease remains hidden. This constitutes a major and long-standing constraint on the articulation of effective policies and programs. Therefore, improving the health of the poor through addressing the problem at the source that bring about profound inequities in health needs specialized tools to view and analyze the data. Following that, data mining was applied for summarizing large volume adults' related problems using BRHP dataset.

The aim of the research is therefore to construct adult mortality predictive model using data mining techniques so as to identify and improve adult health status using BRHP open cohort database. Identifying major determinants and risk factors for adult death helps to alleviate adult mortality problem and helps to limit the loss of the productive group. This in turn calls planners and policy makers to do advocacy efforts on prevention of premature adult death and routine monitoring of adult mortality.

To this end, this research will attempt to explore and answer the following questions.

1. What are the major attributes to consider in applying data mining for adult mortality prediction?

2. Which data mining technique is more appropriate to construct adult mortality predictive model that can be used in adult mortality prediction?

3. What are the optimal variables and determinant factors that lead to adult death in the area of Butajira district?

Methods/Data Mining Modeling

The study followed hybrid methodology of Knowledge Discovery Process (KDP) to achieve the goal of building predictive model using data mining techniques. Hybrid process model was selected since it combines best features of CRISP-DM and KDD methodology to identify and describe several explicit feedback loops which are helpful in attaining the research objectives. Hybrid methodology basically involves six steps (Figure 1).

The Weka GUI chooser

The Weka GUI chooser provides a starting point for launching Weka's main GUI applications and supporting tools. It includes access to the four Weka's main applications: Explorer, Experimenter, Knowledge Flow and Simple CLI.

Methods of training and testing

The classifiers were evaluated by cross-validation using the number of folds. K-fold is a natural number used to check the performance of the model through k–times. It is also suggested that in k-fold algorithm, '10' is about the right number of folds to get the best estimate of error [9]. In each iteration, one fold is used for testing and the other n-1 folds are used for training the classifier. K-folds minimize the bias effects by random sampling of the training and holdout data samples through repeating the experiments ten times. To meet the intention of the research work, 70% was used for training purpose and the remaining 30% for validation of classifier accuracy.

Methods of analysis and evaluation of system performance

After accomplishing model creation, comparing predictive accuracy of the classifiers for unknown tuples is often helpful to evaluate the performance of predictive modeling. It tells us how frequently instances of particular classes are correctly classified as actual class or misclassified as some other classes wrongly.

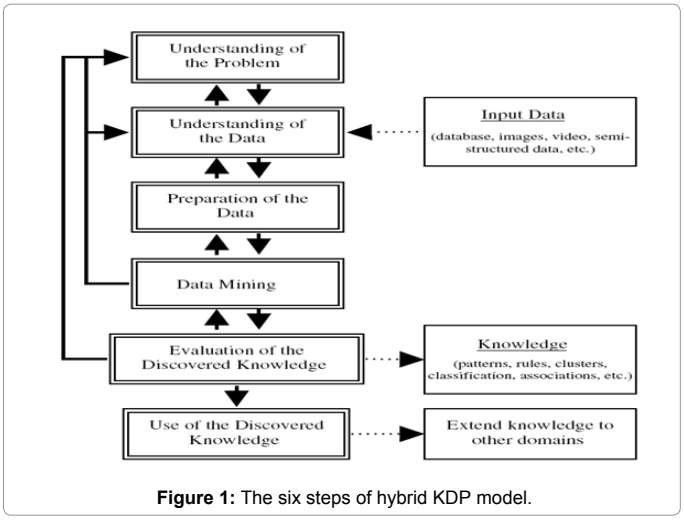


Figure 1: The six steps of hybrid KDP model.

Confusion matrix

Confusion matrix is useful tool for analyzing how well classifier recognized the classes. It is body of table with m by m (row and column) matrix. The row corresponds to correct classification and the column corresponds to the predicted classifications. An entry, CMi,j in the first m rows and m columns indicate the number of tuples of class that were labeled by the classifier as class j [10]. In confusion matrix, there are classifier evaluation metrics like accuracy, error rate, sensitivity and specificity, precision, recall, and F-measure.

Receiver operating characteristic curve

Receiver Operating Characteristic Curve to test which classifier is highly significant for a given subject is determined by ROC analysis and it is becoming widely used tool in medical tests evaluation [11]. This procedure is a useful way to evaluate the performance of classification schemes. For example, it can be used to classify those who are alive and died correctly based on their previous history.

Understanding and pre-processing Butajira rural health programme dataset

The raw data descriptions: Demographic surveillance systems have been established during the past 30 years in a number of field research sites in various parts of the developing countries where routine vital registration systems were poorly developed or nonexistent. Butajira Rural Health Programme has been established in 1986 as an epidemiological study to track a limited and common set of key variables that deal about population dynamics and demographic trends. It is a set of field and computing operations to handle the longitudinal follow up of well-defined entities (individuals, households, and residential units) and all related demographic and health outcomes within a clearly circumscribed geographic area [12].

This epidemiological surveillance data was collected from the nine randomly selected rural kebeles and one urban kebele by implementing probability-proportional to size technique. Events registered by the BRHP are birth, death, marriage, new household, out-migration, in-migration, and internal move (migration within the BRHP surveillance

villages) [13,14]. National and international publications and scientific conferences are the main routes of dissemination of information.

Data selection: BRHP dataset was organized by collecting facts from households located at Butajira district from 1986-2004 in line with epidemiological surveillance in the rural Ethiopia. The dataset contain very large number of variables and instances. The number of instances has been collected after worthy business understanding is taken. From the whole dataset, only population age from 15-60 years were collected in order to satisfy the research goal. Sample data was drawn from both classes using stratified random sampling technique. Sample of records were obtained after stratification for ten peasant associations in a district. This was done by selecting the sample of cases from each peasant association without replacement technique by using statistical software. To select proportional sample of died and alive adults from each peasant association, all the ten villages were stratified into ten strata. From each stratum, using SPSS statistical tool, 25% sample records of 2,715 died and 41,149 alive adults out of 224801 instances were selected randomly. As a result, a total of 43,864 selected cases from the ten villages based on both classes were obtained and used to build and test a model. However, the statistical summaries of attributes relevant to the data mining objectives are performed on the total of 224801 records.

Data preparation: Data preparation generates a dataset smaller than the original one; which can significantly improve efficiency of data mining. This task includes: attribute selection, filling the missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision. Therefore, to increase mining process of the algorithms, understanding the information enfolded in the data was undertaken.

Attributes selection: This is to get a minimum set of best attributes for classification. Related attributes during problem understanding were selected for predictive model building. This is because of considering the effects of predictors with outcome variable whether they are linearly associated. Dataset contains a number of attributes of which many are irrelevant to the mining task and the best and worthy features using tests of statistical significance were selected. Thus, attributes were checked through multi-colinearity analysis for their interdependences. This was checked by using the tolerance and variance inflation factor. Some of the variables that removed are house number, id number, name, mother id, date of death, date of birth, religion, latitude, longitude, father id, tier (serial number of episode), and cause attribute. After having selection of attributes from the entire dataset, statistical summary was done to check their significance effect on the outcome variable using survival Cox regression analysis.

Data decoding: Data decoding was done through allocating a new set of replacement value for a set of values in a given attribute such that each old value can be identified with one of the new values. For the purpose of limiting the confusion during model creation such types of numeric codes were converted to their respective symbols. The alphanumeric code given for the peasant association for nine rural kebeles and one urban kebele was regrouped in to their actual values.

Attributes transformation: Data transformation has momentous effect on data mining since it helps to fix the problem of missing values in the data and brings information to the surface by creating new features to represent trends and other ratios [9]. In this study, through data transformation, age, residence, immigration and outmigration attributes were obtained.

Exploratory data analysis: The attribute's description, data type, unit of measurement, use of frequency tables for the selected attributes was done on the dataset before experimental analysis using data mining tools except some of the derived attributes. Thus, with the use of frequency tables, the exploratory data analysis was performed to detect attributes with the missing values and wrong entries and interpretation of the statistical summary was presented briefly under each of the variable.

Residence: It is code of peasants' village which is string in type with nominal measurement. It was created from the original dataset of peasant association code attribute. In the existing dataset there is PA attribute which is represented by alphanumeric number codes. These codes were regrouped in to two areas, i.e. the nine rural kebles as a rural and the K04 as urban kebele. Majority of cases in residence variable are from Butajira area (27.3% from the total of 224801) while the least number of cases were from the Bido (4.8% from 224801). From all the cases in 22 years follow up, adult mortality in Dobena observed most often (16.8% from 5423 than others while Bido accounts a minimal (5.7% from 5423). In residence variable there are no missing or wrongly entered values (Table 1).

Environmental: This attribute is string in type and nominal measurement. It implies geographical proximity of the district where each episode is came from. It contains different valid distinct values. These are highland, lowland and urban. Majorities of episodes are from highland followed by urban from where the lowest cases involved in the program; 38.7% and 27.2% from the total of 224801 cases in the age group 15-60 respectively. When we see the event and censor in environment variable, from the total cases of death most of them are in lowland (46.4% out of 5423 died cases) followed by urban where the minimum number of death happened (10.1% out of 5423 died cases). Thus, adults living in the lowlands had higher mortality rates (46.4%) than those living in the rural highlands (43.5%) and in urban areas (10.1%) (Table 2).

Sex: It is string in type and nominal measurement. The sex attribute statistics comprised that 51.9% of male adults died during the observation period while female adult death accounts 48.1%. Thus, males died at a slightly higher rate than females and concluded that male sex was found an important associated factor with adult mortality (Table 3).

When comparing sex with area they live, it shows that the risk of dying was higher among rural lowland and highland men and women than urban residents. The sex differential in mortality was most pronounced in the rural lowlands i.e. 46.8% and 46% for female and male respectively (Table 4).

Relation: This attribute is string in type and nominal measurement and it has different distinct valid values. These are head of household, spouse of head, child of head/spouse, parent of head/spouse, other relative and non-relative. Majority i.e. 50.2% (n=2721) of death is in child of head/spouse while the least death was among non-relative and this accounts 0.3% (n=16) (Table 5).

Marital: Marital status is string in type and nominal measurement. In marital attribute, most cases were single; it accounts 33.3% followed by the least valid distinct divorced (0.7%). The finding suggests that more than half of the population death 52.5% (n=2845) were observed in population those reported as too young for marital status and the probability of dying was minimal 0.5% (n=29) in separated group than others. When single marital status is compared with divorced, monogamous, polygamous and widowed, it is the second most factor of adult mortality next to monogamous. In marital status variable there is no missing value observed (Table 6).

Residence				
Residence	Alive	%	Died	%
Mmeskan	17192	7.8	475	8.8
Bati	19967	9.1	653	12.0
Dobena	21035	9.6	911	16.8
Bido	10576	4.8	310	5.7
Dirama	12653	5.8	375	6.9
Yeteker	20656	9.4	572	10.5
Wrib	23381	10.7	627	11.6
Mjarda	14488	6.6	428	7.9
Hobe	18548	8.5	521	9.6
Butajira area	60882	27.8	551	10.2
Total	219378	100.0	5423	100.0

Table 1: Statistical summary of residence attribute.

Environment					
		Alive	%	Died	%
Valid	Highland	84614	38.6	2359	43.5
	Lowland	74090	33.8	2514	46.4
	Urban	60674	27.7	550	10.1
	Total	219378	100.0	5423	100.0

Table 2: Statistical summary of environment attribute.

Sex				
	Alive	%	Died	%
Female	108695	49.5	2606	48.1
Male	110683	50.5	2817	51.9
Total	219378	100.0	5423	100.0

Table 3: Statistical summary of sex attribute.

	Environment						
		Highland		Lowland		Urban		Total
		Count	Row N %	Count	Row N %	Count	Row N %	
Sex	Female	1142	43.8%	1219	46.8%	245	9.4%	100%
	Male	1217	43.2%	1295	46.0%	305	10.8%	100%

Table 4: Statistical summary of sex and environment attribute.

Relation					
		Alive	%	Died	%
Valid	CH	88839	40.5	2721	50.2
	GP	501	.2	23	.4
	HE	43086	19.6	749	13.8
	NR	8089	3.7	16	.3
	RE	26582	12.1	183	3.4
	SP	30837	14.1	550	10.1
	UK	21444	9.8	1181	21.8
	Total	219378	100.0	5423	100.0

Table 5: Statistical summary of relation attribute.

Literacy: This attribute is string in character and nominal measurement. Almost more than half i.e. 54% (n= 2928) of the death among adult were in age group of too young to be at school. The lowest death occurred in the respondents who are able to read only (Table 7).

Education: This attribute is string in character and nominal in measurement**.** From education attribute, mortality was found in too young to be at school is 48% (n=2604) followed by unknown which is 31.6% (n=1716). The lowest death found in further education group 0% (n=1) followed by completed secondary school 5% (28%). The result indicate that the higher the education the lower the probability of dying (Table 8).

Source of water: This attribute is string in character and nominal in measurement. It shows how the source of the water can influence the adult health in given community. Utmost death occurrence was found in the population who were engaged in the river 40.9% (n=2217) followed by well or spring unprotected water users 34.3% (n=1862). On the other hand the lowest death was found in other 0.1% (n=4) followed by lake or pond 0.2% (n=11) (Table 9).

Roof: This is type of roof during episode. It is string in character and nominal in measurement. In business area roof material is considered as an indicator of poverty that adult mortality is prevailed by socio economic conditions. More than 80% (n=4364) of death was found in roof condition of thatched (Table 10).

Windows: This attribute is string in character and nominal in measurement. Living in windowed or un-windowed house has an impact effect in adult death. The result shows that 72.7% (n=3945) of death was found in adult who are living in house with no window. The least cause found on those who report unknown 7.3 %(n=395) (Table 11).

House own: This attribute is string in character and nominal measurement. Majority of the cases were living in owned house 82.7% (n=185932) followed by privately rented house 10. 2% (n= 22861). Death among adult is seen those who live in owned house followed by other; 88.1% and 0.5% respectively (Table 12).

Oxen: This attribute is string in character and nominal measurement. Used to estimate economical condition of a given a family. The more the oxen (two or more) the less the probability of adult death 0.5% (n=29). Inversely the chance of dying in those who reported for unknown oxen and none has positive relationship 65.6% and 21.4% respectively (Table 13).

Rooms: Numeric in type with nominal measurement which is number of rooms in a house. Missing values were filed with their mean for all episodes belonging to the same class using statistical tool (SPSS) (Table 14).

Timad: Numeric in type with scale measurement which is number of timad in a given house. Missing values were filed with their mean for all episodes belonging to the same class using statistical tool (Table 14).

Radius: Numeric in type with nominal measurements. It is radius of circular house in metres in a given house. Missing values were filed with their mean for all episodes belonging to the same class using statistical tool (Table14).

Timex: Numeric in type and scale measurement. Days of exposure during episode (>0). This attribute is numeric in character and scale measurement. It is defined when exposure to the risk of the event in question begins for an event to occur. From this, exposure to the risk of death begins immediately after birth, exposure to the risk of divorce begins with entry into a marriage etc. In BRHP, time of exposure is recorded in days. After selection of attributes from the entire data set in order to verify the significance effect of timex attribute, the days were converted in to respective months. This done by computing time from timex variable by dividing it to 30.25. This number is the mean value day in Gregorian calendaring of the year. After converting days in timex variable to moths, all the attributes whether string or numeric in character were checked for their significance effect on the outcome variable (died, alive). In timex attribute no missing values and errors detected.

Age: Age attribute was derived from date of birth and from the

Marital				
	Alive	%	Died	%
Divorced	1593	.7	33	.6
Monogamous	54246	24.7	590	10.9
Single	74401	33.9	459	8.5
Polygamous	6855	3.1	118	2.2
Separated	1990	.9	29	.5
Too young	25957	11.8	2845	52.5
Unknown	49923	22.8	1085	20.0
Widowed	4413	2.0	264	4.9
Total	219378	100.0	5423	100.0

Table 6: Statistical summary of marital attribute.

Literacy				
	Alive	%	Died	%
Illiterate	22710	10.4	493	9.1
Literate	139277	63.5	1456	26.8
Reading only	2049	.9	65	1.2
Too young to be at school	28667	13.1	2928	54.0
Unknown	26675	12.2	481	8.9
Total	219378	100.0	5423	100.0

Table 7: Statistical summary of literacy attribute.

Education				
	Alive	%	Died	%
No formal education	95221	43.4	1003	18.5
Completed primary school	15025	6.8	71	1.3
Completed secondary school	7419	3.4	28	.5
Further education	842	.4	1	.0
Too young to be at school	17295	7.9	2604	48.0
Unknown	83576	38.1	1716	31.6
Total	219378	100.0	5423	100.0

Table 8: Statistical summary of education attribute

Source of water				
	Alive	%	Died	%
LA	637	.3	11	.2
OT	64	.0	4	.1
PI	52978	24.1	475	8.8
RI	75377	34.4	2217	40.9
UK	3480	1.6	163	3.0
WP	28830	13.1	691	12.7
WU	58012	26.4	1862	34.3
Total	219378	100.0	5423	100.0

Table 9: Statistical summary of source of water attribute.

Roof				
	Alive	%	Died	%
CO	52973	24.1	634	11.7
TH	129807	59.2	4364	80.5
UK	36598	16.7	425	7.8
Total	219378	100.0	5423	100.0

Table 10: Statistical summary of roof attribute.

Window				
	Alive	%	Died	%
NO	131574	60.0	3945	72.7
UK	5531	2.5	395	7.3
YE	82273	37.5	1083	20.0
Total	219378	100.0	5423	100.0

Table 11: Statistical summary of window attribute

last census enumeration by using statistical tool (SPSS). According to the objective of the study, individuals ages 15-60 were selected for this research work. This helps to look the problem existence in different age groups and helps to bring a solution depending on the clue from the study. It is observed that premature death is more prominent i.e. accounts 29.7% (n=1611) followed by youth age group 22.7 %(n=22.7%). In general, more than half of the death in the area was found in adolescent age group (50.4%) as compared to other age groups. The least death was seen in old age group i.e >49. Thus, this early death of the adolescent needs planners and policy makers' advocacy effort to mitigate the sorrowful condition of adult health in the rural (Table 15).

Dishop: Numeric in type and scale in measurement which is distance to Butajira from rounded environment. Most of the death of adult in the area is seen who are living more than five kilometers from the health institution.43.4% while the least death is seen in those who are living very far > 20 kilometers which is 0.1%. In general, inaccessible condition of health care in rural is resulting in adult death. This avoidable means of death also calls attention of government and peoples on Alma-Ata declaration of appropriate health care accessibility which influence the health of the adult (Table 16).

Inmigration: This attribute was created from reason for episode starting. Considering as important parameter of reason for episode start inmigration attribute was created.

Outmigration: This attribute was created from reason for end. Considering Outmigration as important parameter in domain area, it was created from reason for end for these both attributes (rstat and rend) have some similar measurements as a distinct values. Using them as they are in the original dataset may lead to overlapping of the features that affect pattern extraction in knowledge mining.

Status: During the epidemiological survey, a total of 224801 episodes were observed. From this, 219378 and 5423 were registered as alive and event respectively. In the population studied, there are about 97.6% of adults age 15-60 years were alive while 2.4% were recorded as died (Figure 2).

Weka understandable format: After all, the next important issue was importing the selected dataset from SPSS document format into Ms-Excel format in order to create Weka understandable format (arff and csv) for experimentation (Table17).

Description of preprocessed and prepared data: Different activities were performed on the dataset with the objective of making it suitable for the data mining algorithms and producing representative model. Very large numbers of instances were removed and large numbers of attributes are removed.

Experimentation, Analysis and Evaluation of Discovered Knowledge

In this study different experiments were conducted using various data mining methods to derive knowledge from preprocessed data to predict unseen episodes of adults. According to the methodology of this study after preparation of the data, the next task is the mining process. According to Larose [15], if the class attribute is imbalance, this condition further need balancing by different techniques. Unless the classes are proportional, the classification will be skewed in dominant classes. Consequently, the new predicted instances will also fall in the dominant classes erroneously unless the classes' proportionality is considered.

Therefore, sample is assumed to belong to a predefined class, as

House own				
	Alive	%	Died	%
Kebele Or government	7161	3.3	153	2.8
Other	2948	1.3	29	.5
Owned	181152	82.6	4780	88.1
Privately Rented	22793	10.4	68	1.3
Unknown	5324	2.4	393	7.2
Total	219378	100.0	5423	100.0

Table 12: Statistical summary of house own.

Oxen				
	Alive	%	Died	%
None	82702	37.7	1163	21.4
Single animal	51446	23.5	671	12.4
Two or more	6740	3.1	29	0.5
Unknown	78490	35.8	3560	65.6
Total	219378	100.0	5423	100.0

Table 13: Statistical summary of oxen attribute.

		Timad	Room	Radius
N	Valid	174736(77.7%)	212560(94.6%)	132804(59.1%)
	Missing	50065(22.3%)	12241(5.4%)	91997(40.9%)
Total		224801(100%)	224801(100%)	224801(100%)
Mean		2.95	1.53	3.40
Median		3.00	1.00	3.00
Mode		2	1	3
Std. Deviation		2.512	1.017	.724
Skewness		18.067	2.705	.160
Std. Error of Skewness		.006	.005	.007

Table 14: Statistical summary of Timad, room and radius attribute.

Age				
Age	Alive	%	Died	%
15-19(1)	37517	17.1	1611	29.7
20-24(2)	43435	19.8	1229	22.7
25-34(3)	71157	32.4	970	17.9
35-49(4)	48746	22.2	903	16.7
50+(5)	18523	8.4	710	13.1
Total	219378	100.0	5423	100.0

Table 15: Statistical summary of age attribute.

Dishop				
Distance from hospital(km)	Alive	%	Died	%
0.4-5.4(1)	80721	36.8	1098	20.2
5.5-10.5(2)	72560	33.1	2351	43.4
10.6-15.6(3)	48259	22.0	1444	26.6
15.7-20.7(4)	17676	8.1	525	9.7
20.8+(5)	162	.1	5	.1
Total	219378	100.0	5423	100.0

Table 16: Statistical summary of dishop attribute.

Parameters	Original dataset	Target dataset		
Fields	34	21		
Total Number Of Records	320,112	62,869		
File Format	SPSS 16.0	.xls	.csv	.arf
Size of Data	84.3MB	12.1MB	3.05MB	4.83MB

Table 17: Summary of original and target datasets.

determined by the class label attribute (alive, die) and the set of sample used for model construction in this research work is training set by using Weka data mining tool. Thus, the classification accuracy of the minority class become increased in the SMOTE technique for certain level i.e. the total of 62,869 cases (41,149 alive and 21,720 died) were provided and the subsequent experimentations were conducted based on this sample dataset (Figure 3).

For experimentation, different algorithms were employed considering different parameters for model building such as pruning, unpruning and testing model performance with selected attributes and all attributes in both training and testing phases (Table 18).

Selecting and Evaluating the Attributes

Attributes selection involves searching through all possible combination of attributes in the data to find subset of attributes work best for prediction. To do this, determining Cf Subset Eval method was used to assign a worth to each subset of attribute by searching ranker style in the Weka. With this regards, attributes selected using best first techniques in Weka are sex, literacy, education, distance form hospital, age, radius, timad, rooms, in migration and out migration by maximum gain ratio using Weka attribute ranking optimal attributes.

Model Building

To build predictive model, 62,869 instances and 21 attributes were used through using both Naïve Bayes classifier and decision tree algorithm. The models generated with all attributes were compared with models generated with selected attributes. Though the performance variations among different k values are minimal, a bit higher performance was observed in 10 fold k-folds with its minimal error rate to classify the instances in wrong classes (Figure 4). J48 pruned tree model, J48 unpruned tree and Naives Bayes are appeared with competent predictive performance for adult mortality. From all the scenarios experimented, all models reveal the better performance in predicting true positive cases or sensitivity (alive); than predictive performance of true negative case or specificity (died). This is the fact due to the model committing a bias to majority class over the minority class (alive and died) respectively. Testing the model to decide which one of the six models constitutes a better model/classifier of the BRHP data is evaluated using ROC analysis. The ROC area performance of the algorithms show that J48 pruned tree algorithm with all attributes and Naïve Bayes with all attributes scored the highest area of 0.983 while the lowest ROC keeps account in J48 unpruned with selected attributes which is 0.96. With regards of ROC area, a model with perfect accuracy will have an area of 1.0 i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicates

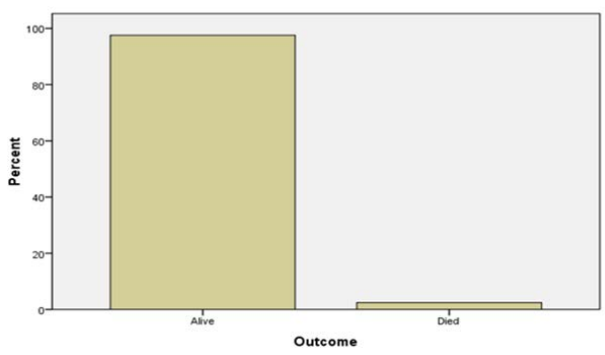
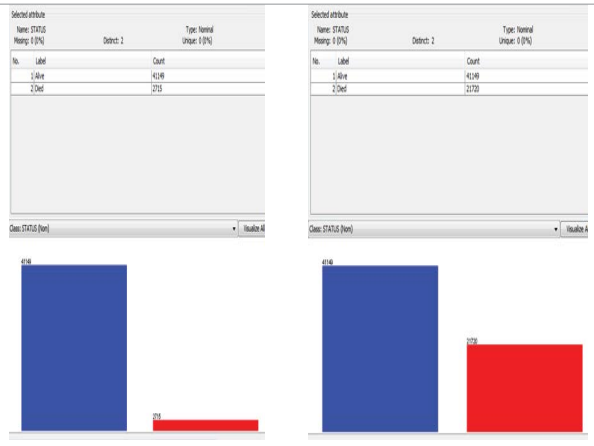


Figure 2: Comparison between died and alive.



(a) Before SMOTE (b) After SMOTE

Figure 3: Side by side review of the class variable using SMOTE.

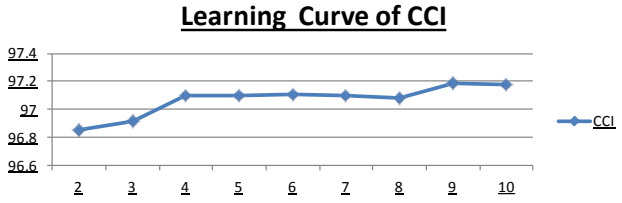


Figure 4: Learning progress curve for different K-values.

the stronger evidence for a positive actual state. Therefore, from the result, J48 pruned tree model and Naïve Bayes classifier have a good capability of 98.3% in adult mortality prediction. Based on the experimentation, J48 pruned tree model with all attributes has an outstanding performance of 97.2% accurate prediction with 0.983 ROC areas (Table 19). And the performance of the models was compared using performance evaluating matrix (Figure 5).

Selected model performance and evaluations

The performance of J48 pruned tree classifier with all attributes gives valuable information in predicting adult mortality as compared to models with all attributes and selected attributes (Figure 6). Thus, J48 pruned tree model with all attributes was selected as the best model and its ROC area was experimented. Initially it moves sharply up from zero showing that the model is better in detecting true positive than false positive. At the end, the curve trade off and become more horizontal showing that from the point where the curve starts to bend to onwards, false positivity outweighs true positivity i.e. the more the curve bend to the right, the more the false positivity rate and the less true positivity rate. The area under the model J48 pruned tree model is 0.983 which is closer to 1 showing that the class value yes gives ROC accuracy of 98.3% (Figure 7).

Result

The predictive performance of J48 pruned tree predictive model was 97.2%. 61,1095 instances were classified correctly while 1774 (2.9%) instances were wrongly misclassified to other class. The model classified 40540 instances as alive out of 41149 instances that in fact they are alive as tested on the test data or which are classified correctly in the class of alive. The remaining 609 instances were misclassified to another class as died despite they are alive.

The model classified 20555 instances as died out of 21720 instances that in fact died and wrongly classified 1165 instances to other class as alive while actually they had died. The model has a good performance in classifying the instances in True class (TP) than True negative class (TN) (alive and died) with predictive performance of 98.5% and 94.6% respectively. Thus, it is possible to conclude that the model is in a good performance to classify True positive than True negative.

From the precision score of the model, the precision of this model for 'Alive' class is a bit higher than precision of 'Died' class (0.972 and 0.971) respectively. With an average precision of 97.2%, instances labeled as belonging for each class Yes and class No (alive, died). From harmonic mean of precision and recall which is F-score, with value of 0.972, it can be concluded that the precision and recall of the model are significantly balanced.

From ROC curve of the selected model, the true positive case (sensitivity) and false positive case (specificity) are represented by vertical axis and horizontal axis, i.e. instances are predicted as alive actually they are alive and predicted as died actually they were died respectively.

Rule extraction

To make a decision tree model more human-readable each path from root to leaf can be transformed into an IF-THEN rule. If the condition is satisfied, the conclusion follows. The algorithm decision tree is the best known method for deriving rules from classification trees. This is simply by traversing any given path from the root node to any leaf. The numbers in (parentheses) at the end of each leaf indicates the number of examples in the leaves. The number of misclassified examples would also be given, after a slash and hence it is possible to compute the success fraction (ratio) to estimate the level of confidence or likelihood of predictability of the class that tells how much the rule is strong.

From the entire models that were generated, J48 pruned tree model with all attributes is selected as the best model for rule generation. This is due to the rules provided by decision tree models can be easily assimilated by human without any difficulty. J48 pruned tree model with all attributes produced different rules. However, the researcher selected best rules that cover most of the data points in the study. After the rule extraction, the researcher turns back to domain experts to discuss up on the generated rules. Some of the rules generated by J48 pruned tree model with all attributes are:

RULE 1: IF *EDUCATION = NO AND TIME OF EXPOSURE <= 1377.333662 AND AGE <= 43 AND OUTMIGRATION = NO AND INMIGRATION = NO AND LITERACY = LI AND WINDOWS = NO AND OXEN = UK AND SEX = M AND ENVIRNMENT = H:* ***THEN*** *the class is Died (119.0/2.0).*

The first rule selected from the rules generated by J48 pruned tree model gives correct result of 119 out of 121 instances that it covers. From this, the likelihood of predictability of the individual to die or death attributed by the above predictors is about 98.4%.

RULE 2: IF *EDUCATION = NO AND TIME OF EXPOSURE <= 1377.333662 AND AGE <= 43 AND OUTMIGRATION = NO AND INMIGRATION = NO AND LITER = LI AND WINDOWS = NO AND OXEN = UK AND SEX = M AND ENVIRONMENT= L:* ***THEN*** *the class is Died (38.0/1.0).*

The second rule selected from J48 tree pruned model is also shows that if the predictors in the above rule 2 are fulfilled, the chance of the person likely to die is 97.4%. All attributes are being constant in the first and second rules, the environment in which the adult lives matter the condition of adult health pattern and it states that living in rural highland and rural lowland is attributable to adult death. Therefore, J48 pruned tree model of data mining technology reveals that if no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death is 98.4% and 97.4% respectively with concomitant attributes in the above rules.

Experiments(1-3)	Scenarios
J48 Unpruned Tree Model Generation	1. J48 unpruned with all attributes
	2. J48 unpruned with selected attributes
J48 Pruned Tree Model Generation	3. J48 pruned tree model with all attributes
	4. J48 pruned tree model with selected attributes
Naïve Bayes Classifier	5. Naïve Bayes with all attributes
	6. Naïve Bayes with selected attributes

Table 18: Experiments and scenarios.

Experiments of the detailed accuracy by class									
Model	Accuracy	TP rate	TN rate	precision	Recall	F-measure	ROC area	size of tree	# of leaves
Scenario1	96.9 %	97.8%	95.2%	0.969	0.969	0.969	0.973	3984	3046
Scenario2	93.6 %	95.6%	89.8%	0.936	0.936	0.936	0.96	3012	1573
Scenario3	97.2%	9 8 . 5 %	94.6%	0.972	0.972	0.972	0.983	1031	715
Scenario4	93.7 %	96%	89.3%	0.937	0.937	0.937	0.964	2043	1054
Scenario5	95.7%	98.1%	91.3%	0.957	0.957	0.957	0.983	-	-
Scenario6	94.7%	97.1%	90%	0.946	0.947	0.946	0.976	-	-

Table 19: Performance summary of the models.

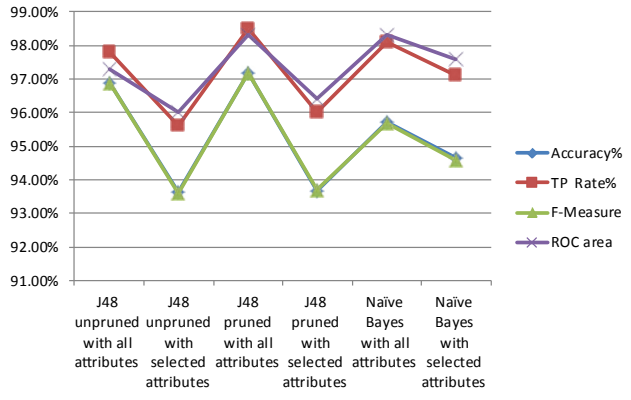


Figure 5: Performance comparison of the models.

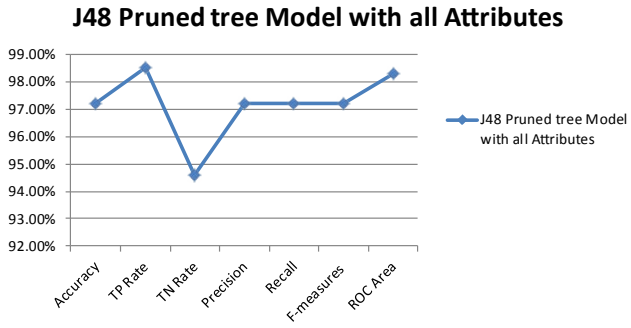


Figure 6: Performace measures of the selected model.

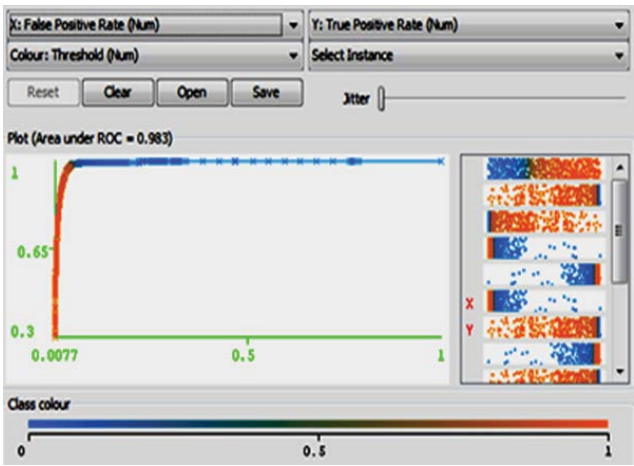


Figure 7: ROC Area of the selected model.

RULE 3: IF *EDUCATION=COMPLETED PRIMARY:* ***THEN*** *the class is Alive(2855.0/33.0)*

RULE 4: IF *EDUCATION = COMPLETED SECONDARY:* ***THEN*** *the class is Alive (1446.0/15.0)*

RULE 5: IF *EDUCATION = FURTHER EDUCATION:* ***THEN*** *the class is Alive (153.0).*

Rule 3, 4 and 5 state that education alone matter the fate of the adult to live or to die without incorporating with other variables. As it has seen in the above rules, the probability of adult to survive will increases as the education attainment of the adult increases. Results from J48 pruned tree model of data mining algorithm reveals that the likely chance of adult to survive in completed primary school, completed secondary school, and further education is (98.9%, 99%, 100%) respectively. This can be explained as, the more the education the more the chance not to die in unfinished old age. The domain experts accepted this rule as education is a central role and mortality can be explained by education and adult with no formal education is more likely to die than adult with education.

RULE 6: IF *EDUCA = UK AND OUTMIGRATION = NO AND INMIGRATION = YES AND LITER = LI AND RESIDENCE = Butajira area 04:* ***THEN*** *the class Alive (383.0/4.0).*

The rule gave a correct result for 383 of the 387 instances that it covers; thus its success fraction is 383/387(98.9%). This rule states that the likelihood of adult to live is 99% with concomitant attributes in

rule 6. This rule is accepted as the residence in which the adult live determines adults' health condition. To this end, from the above rules 1 to 6, the J48 pruned tree algorithm with all attributes suggested that educations plays a considerable role as a root cause of adult death. Adult mortality is also associated with time of exposure, age, outmigration, inmigration, literacy, windows, oxen, environment and sex with the combination of education.

Therefore, J48 pruned tree model could predict status of adult (alive, die) i.e. 97.2% accurate prediction with the respective concordant (True positive and True Negative) with the lower mean absolute error (0.0412) which measures the error between actual and predicted value and with high kappa statistic measures (0.9372); it is usually 1.0 which implies complete agreement. The first six most important parameters/ attributes that determine adult mortality are education status, outmigration, immigration, literacy, residence, and distance from the Butajira town.

Error rate of the selected model

Though the predictive performance of the selected model is promising 97.2% of accuracy for adult mortality prediction, the model commits 2.8% of the cases to classify wrongly to some other classes. The learning algorithm made bias to the majority class. In all the models, the predictive performance in identifying true positive or alive cases is higher than identifying true negative or dies cases. This is because there is imbalance between the two classes in the dataset. Consequently, the model tends to misclassify instances to some other class. The other reason for misclassification is due to the fact that adult mortality status (alive or die) is based on the values of other attributes i.e. taking the similarity of the other attributes as a predominant predictive values.

Conclusion and Recommendation

Conclusion

This study attempted to explore data mining technology on adult predictive modeling using BRHP dataset that can help health care providers in the district to identify adults who are at risk for certain factors. The hybrid, iterative methodology, was employed in this study which consists of six basic steps. In order to generate interesting rule from the huge and massive data collected in the BRHP, a total of 43,864 instances in age group 15-60 years were taken using stratified simple random technique from each peasant association in respective of both classes(died and alive). Knowledge discovery in dataset was employed after having SMOTE technique has been done. The selected variables that used for knowledge mining are residence, environment, relation, sex, marital status, time of exposure, literacy, education, source of water, type of roof, windows in house, house own, oxen, distance from hospital, age, radius, timad, rooms, in migration, and outmigration with status which is the outcome variable.

All the selected attributes were used in the analysis using both decision tree and Naïve Bayes algorithms. Several models were built during experimentation that can predict the risk of adult mortality. Among the models, J48 pruned tree model with all attributes shows an interesting predictive accuracy resulting 97.2% and 98.5% correctly predictive performance of individual as alive cases indeed they are alive.

In summary, it is concluded that the experiments presented in this study show that mortality can be reduced substantially by intervening in certain socio-economic and demographic effects so that probability of adult loose can be minimized. In formulating health policies, the people living at the rural of the socioeconomic strata of the society should get more importance in utilizing the education facilities to reduce avoidable mortality. Other differentiated predictors also need emphasis to come up adult mortality in the district. Enhancing data mining technology particularly the decision tree technique is well applicable to predict adult mortality patterns.

Recommendations

Based on the result of the research finding, the researcher would like to make the following recommendations. This will enhance applicability of data mining technology in adult health prevention and control activities in inline with advocacy efforts of adult mortality reduction policy in rural communities of the country. Thus, the following recommendations could be made considering as they are important issues for further research directions in adult mortality reduction strategies.

- The present study has considered epidemiological dataset to apply data mining in adult mortality prediction. Clinical data that have been gathered from different health care institutions should pay attention in adult mortality reduction. So that future study needs to discover knowledge and patters in clinical datasets and compare it with the result obtained using epidemiological datasets (BRHP).

- Although both the decision tree and Naives Bayes approaches resulted in an encouraging output, still performance improvement is expected. Thus, other classification algorithms such as neural networks and Bayesian network (Belief network) should be tested in order to investigate their applicability to the problem domain in the program by using the entire dataset.

- Although both decision tree and Naïve Bayes reported promising results and hence could be applied in the area of adult mortality predictive modeling, decision tree tends to perform better. Thus, it would be more optimal for the Butajira Rural Health Program to employ the model developed with this technique.

- Results found from this research should be given attention so as to have a better decision making in the Butajira Rural Health Program particularly the program should give special attention to best attribute selected as mortality predictors such as education, time of exposure residence, outmigration, immigration and literacy.

- The health system should focus on healthcare setting health education for the public on focus of attributes identified as adult mortality predictors in the study. Thus, unfinished adult death by family or society due to early death of adult members can be minimized in formulating awareness creation for the people living at the rural area of the socioeconomic strata of the society through giving due emphasis for educational attainment in the area.

- The possibility of incorporating the findings of this study with knowledge based system should be explored so that experts can consult the system in their problem solving and decision making process.

Competing Interests

The authors declared that they have no competing interests.

Acknowledgment

Our earnest gratitude goes Health and Medical sciences college, Addis Ababa University for proper review and approval of this paper. We would also like to extend our gratitude to data collectors for their patience to bring this meaningful information. Our special thanks also extended to Addis Ababa University and Arba Minch College of health sciences for financial support for this study.

Saikia N, Ram F (2010) Determinants of Adult Mortality in India. Asian Population Studies 6: 153-171.

1. Patel I, Chang J, Srivastava J, Balkrishnan R (2011) Mortality in The Developing World Can Pharmacists Intervene? Indian Journal of Pharmacy Practice 4: 2-4.
2. Yamauchi F, Buthelezi T, Velia M (2008) Impact of Prime Age Adult Mortality on Labor Supply: Evidence form Adolescent and Women in South Africa, USA: Washington DC: International Food Policy Research Institute, IFPRI.
3. World Health Organization (2010) World Health Statistics: Geneva, Switzerland WHO Press, Avenue Appia.
4. Mesganaw F (2008) Mortality and Survival from Childhood to Old Age in Rural Ethiopia. Umeå University Medical Dissertations, Sweden, Umeå University, SE-901 87 Umeå.
5. Ngom P, Clark S (2003) Adult Mortality in the Era of HIV/AIDS: Sub-Saharan Africa: Training Workshop on HIV/AIDS and Adult Mortality in Developing Countries. New York: Kenneth Hill, USA.
6. Beaglehole R, Bonita R, Robinson E, Kjellstrom T (1992) The development and evaluation of Basic Epidemiology: Student's Text. Med Educ 26: 482-487.
7. Han J, Kamber M (2001) Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). New York, USA.
8. Berry MJA, Linoff GS (2004) Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 2nd (Edn). John Wiley & Sons, USA.
9. Mehamed Kantardzic JB (2003) Data Mining-Concepts, Models, Methods, and Algorithms. USA: John Wiley & Sons Publication Inc.
10. Krzysztof JC, Witold P, Swiniarski RW, Kurgan LA (2007) Data Mining: A Knowledge Discovery Approach. Springer Science Business Media LLC, New York, USA.
11. Yemane B and Peter B. Butajira DSS Ethiopia, Department of Community Health, Faculty of Medicine, Addis Ababa University and Department of Public Health and Clinical Medicine Umeå University, INDEPTH Monograph: Volume 1 Part C.
12. Yemane B, Stig W, Derege K, Anders E, Fikre E, et al. (1999) Establishing an Epidemiological Field Laboratory in Rural Areas Potentials for Public Health Research and Intervention. The Butajira Rural Health Program 1987-1999. Ethiop J Health Dev 13: 1-47.
13. Ngom P, Binka FN, Phillips JF, Pence B, Macleod B (2001) Demographic surveillance and health equity in sub-Saharan Africa. Health Policy Plan 16: 337-344.
14. Larose DT (2005) Discovering Knowledge in Data-An Introduction to Data Mining. New Jersey USA: John Wiley & Sons Inc.